Embedded Neurally Inspired Visual Processing

William Chapman
Sandia National Laboratories
Albuquerque, New Mexico, USA
gwchapm@sandia.gov

Frances S. Chance
Sandia National Laboratories
Albuquerque, New Mexico, USA
fschanc@sandia.gov

Abstract

Typical digital video processing architectures rely on a largely feed-forward mechanism, yet biological systems exploit complex spatiotemporal dynamics at the level of the sensor to enable highly performant and efficient processing. We present two neurally-inspired algorithms which are specifically designed for embedded sense-and-compute of visual information. The first algorithm utilizes a locally-connected layer of neurons to enable detection of small spatiotemporal patterns. We show that intracellular membrane potentials integrate small information signals over time and enable training techniques, while sparse intercellular communication increases sensitivity to predictable motion patterns. Our second algorithm employs a dynamic membrane time constant to enable multiplicative gain control of inputs. When connected in a locally inhibitive network, the dynamic time networks implement locally divisive gain, akin to standard models of normalization in visual processing. The resulting circuit offers a broader dynamic range than weight-based artificial neural networks, and results in higher performance in image recognition tasks with large input ranges. These algorithms share common requirements of sub-threshold operation and local connectivity. sub-threshold operation and local connectivity in both examples. We explore approaches for a hardware implementation in a simple model that efficiently implements both movement detection and dynamic gain control by utilizing intracellular analog-based processing along with local connections. Our findings demonstrate the importance of an integrative approach to algorithm and circuit design for mapping biologically-inspired algorithms to compact circuit implementations.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Hardware** → **Neural systems**; *Sensors and actuators.*

Keywords

Recurrent Neural Networks, Neural Systems, Computer Vision

ACM Reference Format:
William Chapman and Frances S. Chance. 2025. Embedded Neurally Inspired Visual Processing. In *Great Lakes Symposium on VLSI 2025 (GLSVLSI '25), June 30–July 02, 2025, New Orleans, LA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3716368.3735303

GLSVLSI '25, New Orleans, LA, USA

ACM ISBN 979-8-4007-1496-2/25/06
https://doi.org/10.1145/3716368.3735303

1 Introduction

Simple models of sensory processing typically assume linear-rate encoding, where the number of spikes transmitted from a sensor changes linearly with stimulus intensity. This approach is not dissimilar from standard analog-to-digital conversion, which translates stimulus intensity to an ordinal bit-based value. Biological sensors, however, are highly complex systems that include a number of processing steps, such as frequency decomposition in the cochlea, before a physical signal is translated into a spike-based code. Once signals interact with a receptor on a neuron however, additional processing still occurs in the sensory organ. Lateral inhibition, the Weber-Fechner Law, winner-take-all, sub-threshold coupling, and supralinear responses are all examples of additional interactions that occur in local interactions before transmission to the central nervous system. This results in a spike-train which may have little relationship to the simple magnitude of an input stimulus, but overall has increased the information content of the population code. While this approach requires downstream systems to be "trained" to process this complex nonlinearity, we hypothesize are several potential advantages to be gained from incorporating biologically-inspired smart-compression early in the sensory processing stream. Below we discuss these advantages and how their implementation at or near the sensor benefits system performance. However, we find that while artificial neural networks can be trained to these steps in simulation, the nonlinear nature of these processes can result in catastrophic divergence from ideal behavior. Finally, we end with a call to the VLSI community for approaches that allow these emergent non-ideal interactions to be sufficiently captured to enable end-to-end training of neural networks.

2 Lateral Dynamic Gain Control

All visual processing systems require some form of dynamic gain adjustment, which must occur before analog-to-digital conversion. In typical approaches there is a simple gain adjustment, such as min-max scaling or clipping, that occurs post-sensor, but before analog-to-digital conversion. This step is essential, as analog-to-digital (ADC) devices and subsequent digital processing steps have a limited bit-width. Failure to adjust the gain of devices can result in clipping (Figure 1 middle row), while scaling (bottom row) can preserve global changes in brightness. However, global dynamic gain can result in *catastrophic* information loss in scenarios where the image brightness varies by orders of only in subregions of the field of view.

Neurally-inspired approaches offer a possible alternative approach, based on early local lateral interactions [8, 9]. While standard ANN approaches rely solely on additive interactions, the multiplicative interactions proposed by Heeger can be implemented in neuromorphic approaches. Shunting inhibition, for example, can selectively increase the leak-rate of analog neurons or dendritic

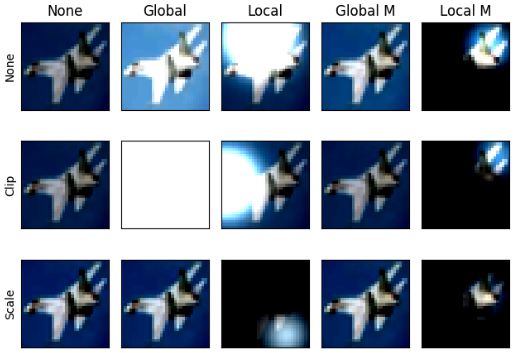

Figure 1: Example of image manipulations that require global or local dynamic gain modulation. The top-left is an example from the CIFAR10 dataset, which we add brightness to either additively or multiplicatively, and either globally or locally. The following rows illustrate an ADC approach that clips to a static range, or implements dynamic gain scaling.

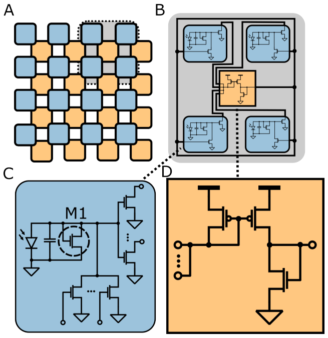

Figure 2: An example of lateral gain control. (A) For a given sensor array, a computational kernel can be applied for a spatial sliding window, allowing for local control of gain, compared to a global system which may result in washout. (B) An example of a 2x2 sliding kernel, which consists of four dynamic sensors (blue) which communicate by a single interaction mechanism (yellow). (C) The dynamic sensor contains a photodiode which charges a local capacitor. The transistor M1 operates in the sub-threshold regime.

compartments to allow an supralinear interaction with additive inputs [2, 4].

Such divisive interactions can mimic the scaling approach of dynamic gain control in a local manner by interacting with a local neighborhood [19], and provide a form of adaptive compression before ADC. A simple candidate implementation is illustrated in Figure 2, which utilizes transistors operating in the sub-threshold regime to regulate leak-rate of photodiode-coupled capacitors (blue circuit). These tunable sensors are then coupled with nearest neighbors through a simple current mirror to allow interactions at a chosen spatial scale. Additional behaviors, such as the strength of coupling, rate of response, and range of sensor output, can be modified by tuning device properties. Critically however, this approach relies on in-sensor processing; if the local shunting interactions occur after digital conversion, then the underlying information has already been lost. While the sub-threshold interactions required for shunting inhibition may vary across devices, some stabilization is provided by recurrent inhibition. Any remaining non-idealities, such as variation of the field of view due to device variability or temporal fluctuations due to transient currents, could be resolved by later layers, as discussed in Section 5.

3 Smart Sensors for Sensitivity

Many tasks involve stimuli which evolve in both space and time, such as tracking objects in a video or identifying a scene based on the interaction of actors. Such tasks can sometimes be processed in a sequential spatial-then-temporal approach, by extracting large scale spatial feature information and evaluating how those features evolve through time. However, in other cases the temporal information may be of higher importance and finer spatial scale. For example, when identifying an object from a distance, integrating temporal information may allow one to detect changes when the spatial resolution was too low to otherwise identify an object [15]. In such cases recurrent processing must occur early in the hierarchy, in order to avoid smoothing small temporal signals before they can be extracted. Recurrent processing is ubiquitous in biological sensory systems, occurring as early as the photoreceptor layer visual processing [16]. In lower-level circuits, hierarchically recurrent processing can increase sensitivity to otherwise weak signals [11, 23]. Here we investigate a particular case for recurrent spatiotemporal processing at early layers of a machine learning model, taking inspiration from biological systems in order to train on binary activations, similar to spiking neurons.

While several use cases may seek to utilize recurrent neural networks for processing of information near sensors, such hardware is often restricted by size, weight, and power (SWaP) constraints. Binarized activation neural networks (BANNs) can minimize the precision of analog-to-digital converters, in the case of physical accelerators such as memristor crossbars [24], or otherwise minimizing the number of binary operations required for linear arithmetic. However, many use cases have temporal dynamics, which requires the use of recurrent neural networks, which requires the storage of state and output over time steps. However, storing these stateful variables and moving them between memory and compute regions is energetically expensive. We therefore require a network that incorporates stateful recurrence, but also minimizes the amount of state that must be retained. This can be achieved either by decreasing the number of layers with statefulness, or by decreasing the amount of state stored by each layer.

Our published work [3] illustrates that first-layer spatiotemporal processing, in the form of a CRNN (Figure 3) is essential for enabling low signal-to-noise tracking. In real-valued networks a first-layer CRNN is not essential, but in spiking neural networks, training fails if the first layer is not CRNN. We rationalize that first-stage spiking CRNNs have a unique access to raw sensory signals, and can integrate sub-decision (spiking) evidence over multiple frames. In contrast, a stateless binary activation layer cannot integrate such evidence, increasing the expressive power compared to binary activation functions.

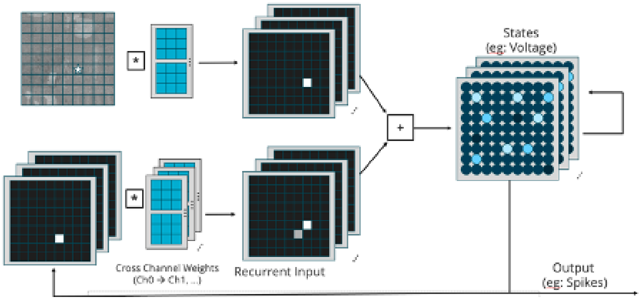

Figure 3: Example of a convolutional recurrent network. The top path is identical to the pathway of a standard convolutional network, in which input images (here one channel) are transformed by multiple kernels to create a multichannel output. The bottom path is similar, but is now applied to the output activity of the layer on each timestep. Both pathways are added to create an update of the local state, allowing spatial information from the input stream to interact with simple local recurrent processing before undergoing processes such as digitization.

Recurrent activity alone is also insufficient to enable such binary activation networks. This can be intuitively understood from Figure 4, which illustrates activity and and gradients in a recurrent neural network trained on a temporal task. For floating point activations, gradients are smooth through time, while binary activations result in a binary and sparse gradient, which is recovered with a spiking (LIF-based) neural network. This illustration highlights the need for hardware time-constants of sufficient scale to span multiple "frames" in order to smooth activity and allow training of binary networks.

4 Higher Order Feedback Dependent Processing

Neocortically inspired circuits, despite their central nervous system origin, offer additional insights for in-sensor computing. While the retinally-inspired circuits above focus on sensory-driven activity with local interactions, cortical circuits routinely integrate abstracted (or 'contextual') feedback information to modulate feedforward processing. Such approaches could be used in combination with in-sensor computations to increase early-stage processing in response to high-level information that may be extracted in more a more complex system, such as a large digital neural network. For example, early-stage lateral interactions can induce direction-specific increases in sensitivity to specific motion directions, but are limited to specific scales or patterns of motion [18]. Through the use of hierarchical feedback however, these motion primitives may be dynamically expanded, to perform a form of Bayesian inference [1], while higher-order systems are responsible for determining the candidate kernels from recent observations. For scenarios such as low-SNR tracking, where online integration of attention-modulated signals is necessary for initial detection, this approach may allow more general sensitivity.

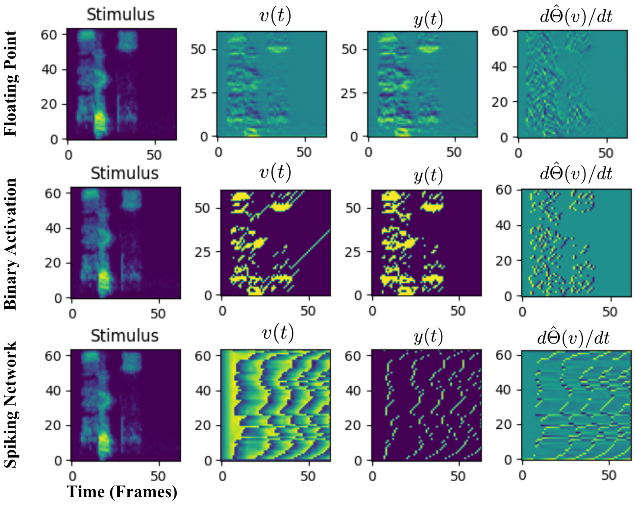

Figure 4: Illustration of a convolutional-recurrent network performing a temporal classification task. The stimulus is presented column-by-column resulting in feedforward activities v(t), and output from the first layer of the network y(t). Partial gradients with respect to feedforward activity are shown in the final column. For real-valued networks the gradients are smooth, while binary activation networks are discrete valued and discontinuous. Spiking networks have discontinuities at spike times, but are largely smooth over time, due to the intrinsic integration time within individual neurons.

Despite the computational complexity of such canonical circuits, the underlying neuronal compartments are relatively simple, consisting of passive dendritic compartments and active (possibly adaptive) spiking somatic compartments. Recent work has demonstrated compact circuit implementations [17]. Even simpler mechanisms of context-dependent modulation may occur through time-constant modulation [6], voltage thresholds for spiking [10], or dynamic switching between intermediate readout circuits.

5 Needs from Neuromorphic

The methods outlined above are examples of neurally inspired computation which have been embedded in analog processing circuits, as opposed to other neuromorphic approaches which utilize saltatory spiking activity and minimal intracellular analog-regime processing. While neural networks, including bio-inspired approaches, are tolerant to some degree of noise and variability in device properties, they are limited. Biological systems address this variability with a variety of on-line learning mechanisms, negative feedback, and homeostatic mechanisms. In electronic systems however, we are largely limited to two approaches: on-device learning or hardware-aware training. On-device learning has the potential to overcome hardware nonidealities, including drift or degradation over time, but also requires specialized hardware capable of implementing supervised learning rules [7, 21] and limits the depth of neural processing to relatively shallow approaches [25].

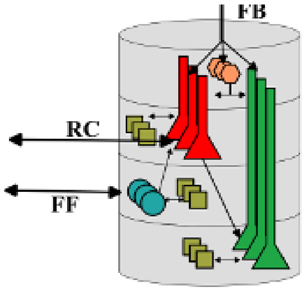

Figure 5: Illustration of a canonical cortical microcircuit. Pyramidal neurons (red/green) are spatially extended, allowing separate input sites for feedforward (FF) and feedback (FB) pathways. The distal dendrites of the pyramidal neurons are able to *modulate* responses, but can not directly elicit action potentials, due to their distance from the soma.

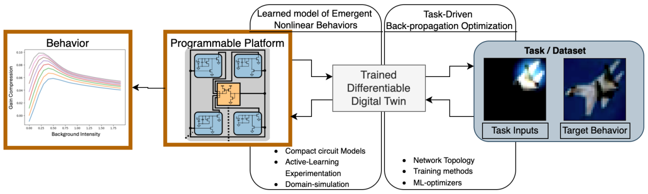

Figure 6: An extended form of hardware aware training. Rather than directly fitting neural networks to specific device behavior and task performance, we suggest a highly accurate and variational digital twin. The underlying behavior of the digital twin can be fit to device behavior, similar to physics aware training or physics-informed neural networks. The accurate twin can then select from candidate parameter spaces to perform tasks of interest.

Hardware-aware training (HAT) shows greater promise as a general purpose mechanism for training arbitrary networks for deployment on analog devices. HAT may be implemented with device-in-the-loop (or "physics aware") training, which utilizes hardware to perform forward inference and digital approximate models for credit assignment [20]. Such approaches have been demonstrated for a variety of devices [13, 14], but scale poorly to mass manufacturing. A more promising approach is to develop sufficiently high-fidelity digital models which allow emulation of device non-idealities and emergent behavior from interactions of devices. Such an approach would be similar to standard approaches in neuromorphic computing, where intrinsic properties of leaky-integrate and fire neurons are modeled and fit with surrogate gradient descent [5], and more recently in models utilize dendritic structures for additional expressivity [22]. Such a model could be used for both forward and backward phases of training, and training an network to a distribution of specimen-specific behaviors, would implement a form of variational training, resulting in a network which is robust to any given final device. Previous work has demonstrated how compact variational models of individual devices can be integrated into high-fidelity models (e.g: SPICE [12]), however the computational needs for such approaches prohibit large-scale simulations that would be necessary for variational training. Currently, the missing link to enable these approaches, and therefore embedded neuromorphic processing, is an approach for granular simulation of systems of devices that can be incorporated into such hardware aware training.

Acknowledgments

The authors acknowledge financial support from the Laboratory Directed Research and Development Program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

References

[1] Andre M Bastos, W Martin Usrey, Rick A Adams, George R Mangun, Pascal Fries, and Karl J Friston. 2012. Canonical Microcircuits for Predictive Coding. *Neuron* 76, 4 (2012), 695–711. doi:10.1016/j.neuron.2012.10.038

[2] Frances S Chance and Suma G Cardwell. 2023. Shunting Inhibition as a Neural-Inspired Mechanism for Multiplication in Neuromorphic Architectures. In *Proceedings of the 2023 Annual Neuro-Inspired Computational Elements Conference*. 41–46.

[3] G. William Chapman, Corinne Teeter, Sapan Agarwal, T. Patrick Xiao, Park Hays, and Srideep S. Musuvathy. 2024. Biological Dynamics Enabling Training of Binary Recurrent Networks. In *2024 Neuro Inspired Computational Elements Conference (NICE)*. 1–7. doi:10.1109/NICE61972.2024.10549632

[4] Jordan Edwards, Luke Parker, Suma G. Cardwell, Frances S. Chance, and Scott Koziol. 2024. Neural-Inspired Dendritic Multiplication Using a Reconfigurable Analog Integrated Circuit. In *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. 1–5. doi:10.1109/ISCAS58744.2024.10557895

[5] Jason K Eshraghian, Max Ward, Emre O Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, and Wei D Lu. 2023. Training Spiking Neural Networks Using Lessons from Deep Learning. *Proc. IEEE* (2023).

[6] Romain Ferrand, Maximilian Baronig, Thomas Limbacher, and Robert Legenstein. 2023. Context-Dependent Computations in Spiking Neural Networks with Apical Modulation. In *International Conference on Artificial Neural Networks*. Springer, 381–392.

[7] Will Greedy, Heng Wei Zhu, Joseph Pemberton, Jack Mellor, and Rui Ponte Costa. 2022. Single-Phase Deep Learning in Cortico-Cortical Networks. *Advances in Neural Information Processing Systems* (2022). doi:10.48550/ARXIV.2206.11769

[8] David J. Heeger. 1992. Normalization of Cell Responses in Cat Striate Cortex. *Visual Neuroscience* 9, 2 (Aug. 1992), 181–197. doi:10.1017/S0952523800009640

[9] David J. Heeger and Klavdia O. Zemlianova. 2020. A Recurrent Circuit Implements Normalization, Simulating the Dynamics of V1 Activity. *Proceedings of the National Academy of Sciences* 117, 36 (Sept. 2020), 22494–22505. doi:10.1073/pnas.2005417117

[10] Peter Helfer, Corinne Teeter, Aaron Hill, Craig M Vineyard, James B Aimone, and Dhireesha Kudithipudi. 2023. Context Modulation Enables Multi-Tasking and Resource Efficiency in Liquid State Machines. In *Proceedings of the 2023 International Conference on Neuromorphic Systems*. 1–9.

[11] Laurent Itti and Pierre Baldi. 2009. Bayesian Surprise Attracts Human Attention. *Vision Research* 49, 10 (June 2009), 1295–1306. doi:10.1016/j.visres.2008.09.007

[12] Paul Kuberry and Eric Keiter. 2021. *An Embedded Python Model Interpreter for Xyce™ (Xyce-PyMi)*. Technical Report SAND2021-9504.

[13] Shuaifeng Li and Xiaoming Mao. 2024. Training All-Mechanical Neural Networks for Task Learning through in Situ Backpropagation. *Nature Communications* 15, 1 (Dec. 2024), 10528. doi:10.1038/s41467-024-54849-z

[14] Jiaqi Lin, Sen Lu, Malyaban Bal, and Abhronil Sengupta. 2024. Benchmarking Spiking Neural Network Learning Methods with Varying Locality.

[15] Tian J Ma and Robert J Anderson. 2023. Remote Sensing Low Signal-to-Noise-Ratio Target Detection Enhancement. *Sensors* 23, 6 (2023), 3314.

[16] Niru Maheswaranathan, Lane T McIntosh, Hidenori Tanaka, Satchel Grant, David B Kastner, Joshua B Melander, Aran Nayebi, Luke E Brezovec, Julia H Wang, Surya Ganguli, et al. 2023. Interpreting the Retinal Neural Code for Natural Scenes: From Computations to Neurons. *Neuron* (2023).

[17] Maryada, Chiara De Luca, Arianna Rubino, Chenxi Wen, Matteo Cartiglia, Ioan-Iustin Fodorut, Melika Payvand, and Giacomo Indiveri. 2025. A Canonical Cortical Electronic Circuit for Neuromorphic Intelligence. 2025.03.28.646019 pages. doi:10.1101/2025.03.28.646019

[18] Alex S. Mauss, Anna Vlasits, Alexander Borst, and Marla Feller. 2017. Visual Circuits for Direction Selectivity. *Annual Review of Neuroscience* 40, Volume 40, 2017 (July 2017), 211–230. doi:10.1146/annurev-neuro-072116-031335

[19] Carver A. Mead and M.A. Mahowald. 1988. A Silicon Model of Early Visual Processing. *Neural Networks* 1, 1 (Jan. 1988), 91–97. doi:10.1016/0893-6080(88)90024-X

[20] Ali Momeni, Babak Rahmani, Matthieu Malléjac, Philipp del Hougne, and Romain Fleury. 2023. Backpropagation-Free Training of Deep Physical Neural Networks. *Science* 382, 6676 (Dec. 2023), 1297–1303. doi:10.1126/science.adi8474

[21] Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A. Richards, and Richard Naud. 2021. Burst-Dependent Synaptic Plasticity Can Coordinate Learning in Hierarchical Circuits. *Nature Neuroscience* 24, 7 (July 2021), 1010–1019. doi:10.1038/s41593-021-00857-x

[22] Mark Plagge, Suma G Cardwell, and Frances S Chance. 2024. Expressive Dendrites in Spiking Networks. In *Neurally Inspired Computing Elements.*

[23] Rajesh P. N. Rao and Dana H. Ballard. 1999. Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects. *Nature Neuroscience* 2, 1 (Jan. 1999), 79–87. doi:10.1038/4580

[24] \relax TP Xiao, \relax WS Wahby, \relax CH Bennett, P Hays, V Agrawal, \relax MJ Marinella, and S Agarwal. 2023. Enabling High-Speed, High-Resolution Space-Based Focal Plane Arrays with Analog in-Memory Computing. In *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 1–2.

[25] Xiaohui Xie and H. Sebastian Seung. 2003. Equivalence of Backpropagation and Contrastive Hebbian Learning in a Layered Network. *Neural Computation* 15, 2 (Feb. 2003), 441–454. doi:10.1162/089976603762552988